nature portfolio

Peer Review File

Large disagreements in estimates of urban land across scales and their implications

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this manuscript, the authors compared several global urban land datasets and their differences at different scales. The authors also discussed the impact of differences of these datasets on urban climate, urban environmental risks, and so on. The workload for this manuscript is still quite substantial. However, there is a lack of innovation and too many uncertainties for observational and modeling applications using the different global urban land products. The conclusions of this manuscript are unclear, and there is nothing new. Some of the results are factual. Therefore, rejected is my opinion towards this manuscript with several reasons given as follows:
(1) As mentioned in your manuscript, due to different spatial resolution, different methods used, and different urban definitions (impervious surface, built-up area, and human settlement and so on), there are great differences between different global urban land products. Many scholars have also conducted comparative studies on different global urban land products. Although you also did a lot of comparison work, what is your innovation? How do your conclusions differ from theirs?
(2) About the data products used in the manuscript, firstly, The GHSL products are also long-term series and relatively commonly used datasets. In addition, in my opinion, GlobeLand30 is also a very good datasets (https://www.un-spider.org/links-and-resources/data-sources/land-cover-map-globeland-30-ngcc). Therefore, this manuscript should compare the two datasets. From my research experience, the accuracy of 10-m resolution global urban land products, for example, ESA and ESRI land cover, is very low. But, in Table S2, their accuracy is still very high. Therefore, I think there is something wrong with your accuracy assessment. In this manuscript, you should give the absolute value of the area, not the percentage of impervious surface. For example, the total areas of global urban land.
(3) While the manuscript undertakes a commendable effort in comparing various global urban land datasets, it appears to miss a crucial opportunity to directly address and enhance the practical application and customization practices already prevalent among urban researchers. For instance, it is common for researchers to adapt datasets like WSF2019 by integrating additional layers such as roads from OpenStreetMap, acknowledging that WSF2019 itself incorporates such data for SVM classification training. Furthermore, the practice of applying definitional adjustments to include urban green spaces or public areas, thereby transforming datasets like ESA and WSF for specific research needs, is well-documented. However, the manuscript does not seem to delve into how its findings could offer new insights or breakthroughs for these existing approaches. The potential of the study to guide researchers in navigating the complexities of dataset customization and to provide actionable insights for handling specific urban research challenges remains unclear. This gap suggests a missed opportunity to contribute to the ongoing discourse on the practical application of urban land datasets, potentially limiting the manuscript's relevance and utility for the broader community of urban researchers looking for novel methodologies and deeper understandings to enhance their data processing and application strategies.

Minor revisions are as follows:
(1) This sentence ‘there are currently at least four ~10 m resolution global land use land cover products, which include urban classes, and several urban-specific datasets that span multiple

decades' is not correct, the earliest 10-meter global urban land product was in 2015.
(2) In the figure 3, I recommend using the absolute value of the area rather than the percentage of impervious surface.
(3) It is recommended to add some references about other global urban land products, for example Hi-GISA (10-m resolution), Globeland30, GLC-FCS30D, and so on.

Reviewer #2 (Remarks to the Author):

The present study investigates a number of global urban land cover datasets and compare them w.r.t. to urban land cover fraction and long-term changes in the urban land cover. The authors have also discussed the implications of the differences between the datasets for practical examples and have produced a web tool. The analysis is comprehensive and well explained. Furthermore, the study covers an important topic in many different communities, especially for the (urban) climate modelling community. In particular, the different definitions of urban areas and the different resolutions of the data sets are the main reasons for the differences. With this analysis, the others provide guidance for different applications of the datasets. However, there are minor issues that need to be addressed before it can be published.
Minor:
I like that the authors show examples on the impact of the differences on different practical applications. However, for one the example of urban land in flood plains there is only one sentence in the result section incl. one subfigure (Figure 4b). It would be interesting to explain the huge differences in Asia and Oceania.
Line 56: It would be good to shortly introduce the selection criteria for the datasets her. Otherwise, the reader is wondering what "almost all datasets" mean.
Lines 69ff: use 1 km2 instead of sq. km
Line 73: Esri Land Cover
Lines 81-83: Please remove "etc." since it is not clear, which countries are meant with "etc."
Lines 86-88: You are talking about urban growth here but the analysis of urban growth comes later in the manuscript. This can confuse the reader.
Lines 91ff: You talk about "global to local scales" and "large scale". However, it is not really clear, what the difference between global and large scale. Also local scale can have a different meaning in different communities. Hence, it would be good to define these scales.
Lines 242-246: This sentence is long and confusing. It is not clear, why there should be an overestimation if not all urban classes are used. It is also not clear, what is overestimated.
Line 253: If roads are removed the assumption for some urban parameterization is not valid. Some assume an urban canyon, with buildings and roads. Hence, the recommendation should avoid such datasets for particular urban parameterizations.
282-300: While I think that future urbanization is an important topic, I am wondering why a new analysis is conducted in the discussion section. One could consider this as an outlook. It could also be listed as another application.
Caption Figure 4b: The sub-figure does not show correlations but scatterplot/linear regressions.
Line 334: "one of the most"

Line 396: Are there also negative values? Maybe you mean that both datasets need to have non-zero values?
Figures 3d and S6: Previous studies also showed that there is a rapid increase rapid urban expansion in Europe in the ESA-CCI LC dataset (e.g. Li et al. 2021, Hoffmann et al. 2023), which is not shown in the other datasets. This increase can be also seen in the global urbanization rates (Li et al. 2018). Is there an explanation for this behavior? In any case, it would be good to mention this behavior in the manuscript.
Web-tool: This is really interesting and helpful. However, I found that for MODIS tilted grids are show e.g. in Texas.

Li, W., MacBean, N., Ciais, P., Defourny, P., et al. (2018): Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015), Earth Syst. Sci. Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-2018
Radwan, T.M., Blackburn, G.A., Whyatt, J.D. et al. (2021) Global land cover trajectories and transitions. Sci Rep 11, 12814. https://doi.org/10.1038/s41598-021-92256-2
Hoffmann P, Reinhart V, Rechid D, de Noblet-Ducoudré N, et al. (2023): High-resolution land use and land cover dataset for regional climate modelling: Historical and future changes in Europe. Earth Syst. Sci. Data, 15, 3819–3852, doi: 10.5194/essd-15-3819-2023

Reviewer #3 (Remarks to the Author):

The manuscript addresses an increasingly important topic. The proliferation of global datasets, enabled by the decreasing computational costs and increasing availability of remotely sensed imagery at higher spatial resolutions and temporal frequencies, brings up the questions of to what extent these datasets indeed represent what they are claimed to represent and the degree of agreement among them. This is also one of the best written manuscripts I have reviewed for a long while. However, I would also like the authors acknowledge another, earlier work on comparing global urban land cover datasets, and relate their approach and findings to those of that study. Specifically, I would like to see a broad comparison and discussion of the another recently published study (Mu et al. 2022) not cited in the current manuscript. I would like to know in what ways the authors work is similar or not to that one with an emphasis on where these two studies agree (or not). Other than that, I have a few suggestions and comments to help improve the clarity and flow of the paper (please see the annotated manuscript and supplementary file).

Reference:
Mu, H., C. Xu, D. Liu, X. Li, Y. Zhou, P. Gong, J. Huang, X. Du, J. Guo, W. Cao and Z. Sun (2022). "Identifying discrepant regions in urban mapping from historical and projected global urban extents." All Earth 34(1): 167–178.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Large disagreements in estimates of urban land across scales and their implications

(Original comments in **bold**, response in regular font, and manuscript revision in green)

We thank the reviewers for the constructive comments on our manuscript. We have significantly revised this version to address all the reviewers' comments, which has substantially improved the clarity and robustness of the conclusions. Of note, we have:

1) Added more discussions on the policy implications of these disagreements on line 420:

Various urban datasets have been and are being used for informing policy and decisions across scales. Urban planners at the municipal and national level rely on maps of urban land and its evolution over time to design policies that ensure sustainable urban development, relevant to Sustainable Development Goal 11 (SDG-11) charted by the United Nations Member States[77]. Although urban decision makers may have some detailed local maps at the city level, these are generally static, leading to reliance on satellite-derived products to explore patterns of urban compaction and sprawl over time and set policies consistent with, for example, the 15-Minute City, 3-30-300, and Compact City paradigms[78–80]. Consequently, many researchers have attempted to generate datasets to aid this sort of decision making, frequently constrained by estimates of urban land; similar to those examined here[81,82]. Urban datasets, as incorporated into process-based models, are also used to quantify future urban climate impacts, such as on extreme heat and precipitation events, to inform government agencies[83]. At the national level, authorities will soon have to report on changes in urban extent and the ecosystem services this affects under the newly adopted statistical standard for the System of Environmental-Economic Accounting Ecosystem Accounting (SEEA EA)[84]. This policy framework aligns with the International Union for Conservation of Nature (IUCN) ecosystem typology[85], which includes a mosaic definition of urban areas, thereby aligning more with products like Dynamic World and others that adopt a larger minimum mapping unit. However, regardless of the urban dataset chosen, ecosystem accounting practitioners are encouraged to adopt design-based methods for estimating areas from maps derived with remote sensing[86]. Finally, explicit future projections of urbanization have started to be incorporated into Integrated Assessment Models (IAMs), which are often used to set national-scale energy policies and emission targets[87]. Given the variations in the temporal and spatial patterns of urban land across datasets seen here, we suggest that decision makers consider a variety of datasets to inform policies from urban to national scales. Similarly, data producers should try to be more up front about the assumptions underlying their datasets, particularly when the intent is to assist policymaking. In a world of cheap compute and increasingly available high-resolution satellite imagery, we expect such global data products to continue to be published. Given this rapid pace of dataset generation, a concentrated and ongoing effort is needed from the urban scientific community to assess fit-for-purpose datasets for distinct science and applications before they are adopted to quantify the costs and benefits of environmental solutions and support broader policies.

2) Added a new analysis of gridwise disagreement to compare our results against that of Mu et al. (2022), which is the only similar study of its kind (also suggested by Reviewer #3) on line 105:

We also calculate the coefficient of variation between the eight present-day estimates for 0.9° x 1.25° grids over the Earth's surface (Fig. 2b). This approach is similar to that of Mu et al. (2022)[26] and avoids the variabilities in how regions should be defined beyond geopolitical boundaries. Although we use a different (partially overlapping) set of urban data products compared to the Mu et al. (2022)[26] study, there are some common regions with disagreements between datasets in both studies, namely Southeast and Central Asia. We also find large disagreements in East Africa (Fig. 2b).

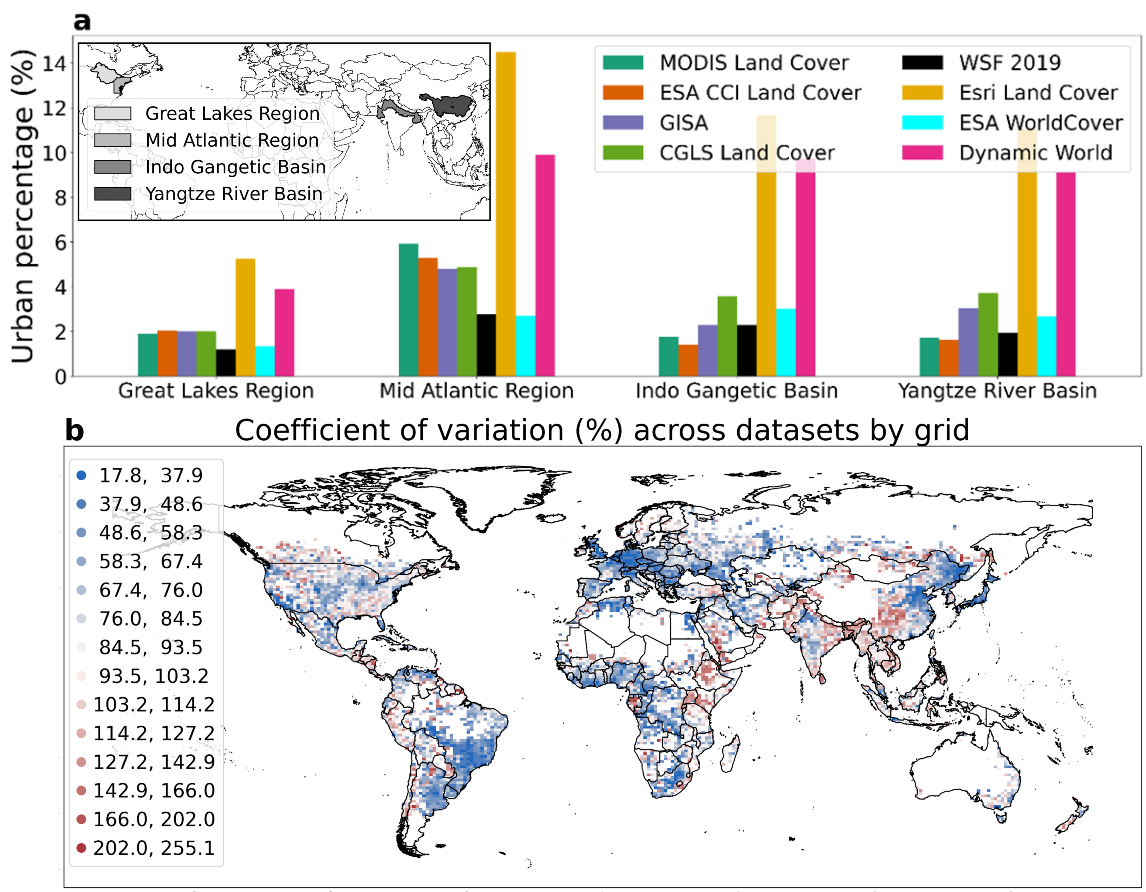


Fig. 2 | Present-day urban estimates across datasets from regional to grid scale | Sub-figure **a** shows the urban percentage across eight datasets for four selected regions in the world. The extent and location of these regions is shown in the inset. Sub-figure **b** shows the coefficient of variation (standard deviation divided by mean) among those urban estimates for 0.9° x 1.25° grids. The legend value ranges exclude the upper bound.

3) We have also expanded the discussion on how the broader scope of this study differs from the primarily remote sensing focuses studies on urban land cover on line 369:

In light of the development of multiple new global urban datasets at finer resolutions in the last decade, our goal here was to examine whether these state-of-the-art products provide better constraints on our understanding of urbanization across scales. The datasets considered here include thirteen historical urban land cover products (plus GHSL; see Methods), two datasets specifically used in process-based urban models, and four future projections of urban land. Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts.

4) Expanded on a few discussions for clarity, including reshuffling some of the figures, to expand on some of the important implications of the study.

The point-by-point responses to all the reviewers' comments are below.

Reviewer #1 (Remarks to the Author):

In this manuscript, the authors compared several global urban land datasets and their differences at different scales. The authors also discussed the impact of differences of these datasets on urban climate, urban environmental risks, and so on. The workload for this manuscript is still quite substantial. However, there is a lack of innovation and too many uncertainties for observational and modeling applications using the different global urban land products. The conclusions of this manuscript are unclear, and there is nothing new. Some of the results are factual. Therefore, rejected is my opinion towards this manuscript with several reasons given as follows:

We thank the reviewers for recognizing the amount of work that went into this study. We should note that one of the reasons for doing this analysis was precisely to focus on how choice of dataset leads to uncertainties in science and applications, which is different from the bulk of the literature on this topic focused on developing new datasets and for accuracy assessments. We have expanded on this in the revised manuscript on line 52:

Previous studies that have explored these discrepancies have focused on earlier-generation datasets that were generally coarser (~1 km), in line with the resolutions of the commonly deployed Earth observing satellites of the time, and did not include enough observations to provide time series of urban expansion[23,24]. More recent comparisons of higher resolution datasets, primarily in the remote sensing literature, are limited because they are either restricted to regional extents[25], or focus on comparing product accuracies, not area estimates and product typologies and their implications[19]. One recent study[26] examined the discrepancies between six 30 m global urban land products, though the main implication of these differences examined was for future urban projections. Here, we provide a comprehensive comparison of almost all (see Methods for selection criteria) the medium to fine resolution (100 m to 10 m) global urban datasets currently available (Table S1), showing that definitions of 'urban' remain a critical issue across these datasets, particularly in the newer 10 m resolution products. More importantly, we examine the consequences of the choice of dataset on a few common use cases relevant for examining urban climate change and its human impacts. For these use cases, the choice of dataset can have large influence on magnitude of urbanization impacts and, in some cases, even the direction of the urban climate signal.

And on line 369:

In light of the development of multiple new global urban datasets at finer resolutions in the last decade, our goal here was to examine whether these state-of-the-art products provide better constraints on our understanding of urbanization across scales. The datasets considered here include thirteen historical urban land cover products (plus GHSL; see Methods), two datasets specifically used in process-based urban models, and four future projections of urban land. Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts.

We absolutely agree with the reviewer that there are uncertainties in these products. But they are often used for both science and applications. We caution against using these products without properly examining if they are fit-for-purpose. We have thus added a new Concluding paragraph in the Discussion section (given the format of nature portfolio journals) that clarify the broader implications, including policy implications, of this line 420:

Various urban datasets have been and are being used for informing policy and decisions across scales. Urban planners at the municipal and national level rely on maps of urban land and its evolution over time to design policies that ensure sustainable urban development, relevant to Sustainable Development Goal 11 (SDG-11) charted by the United Nations Member States[77]. Although urban decision makers may have some detailed local maps at the city level, these are generally static, leading to reliance on satellite-derived products to explore patterns of urban compaction and sprawl over time and set policies consistent with, for example, the 15-Minute City, 3-30-300, and Compact City paradigms[78–80]. Consequently, many researchers have attempted to generate datasets to aid this sort of decision making, frequently constrained by estimates of urban land; similar to those examined here[81,82]. Urban datasets, as incorporated into process-based models, are also used to quantify future urban climate impacts, such as on extreme heat and precipitation events, to inform government agencies[83]. At the national level, authorities will soon have to report on changes in urban extent and the ecosystem services this affects under the newly adopted statistical standard for the System of Environmental-Economic Accounting Ecosystem Accounting (SEEA EA)[84]. This policy framework aligns with the International Union for Conservation of Nature (IUCN) ecosystem typology[85], which includes a mosaic definition of urban areas, thereby aligning more with products like Dynamic World and others that adopt a larger minimum mapping unit. However, regardless of the urban dataset chosen, ecosystem accounting practitioners are encouraged to adopt design-based methods for estimating areas from maps derived with remote sensing[86]. Finally, explicit future projections of urbanization have started to be incorporated into Integrated Assessment Models (IAMs), which are often used to set national-scale energy policies and emission targets[87]. Given the variations in the temporal and spatial patterns of urban land across datasets seen here, we suggest that decision makers consider a variety of datasets to inform policies from urban to national scales. Similarly, data producers should try to be more up front about the assumptions underlying their datasets, particularly when the intent is to assist policymaking. In a world of cheap compute and increasingly available high-resolution satellite imagery, we expect such global data products to continue to be published. Given this rapid pace of dataset generation, a concentrated and ongoing effort is needed from the urban scientific community to assess fit-for-purpose datasets for distinct science and applications before they are adopted to quantify the costs and benefits of environmental solutions and support broader policies.

(1) As mentioned in your manuscript, due to different spatial resolution, different methods used, and different urban definitions (impervious surface, built-up area, and human settlement and so on), there are great differences between different global urban land products. Many scholars have also conducted comparative studies on different global urban

land products. Although you also did a lot of comparison work, what is your innovation? How do your conclusions differ from theirs?

While there have been previous studies on comparing urban land cover products, many of which we have cited in our paper, many of these comparisons are outdated given the pace of dataset generation, especially with the release of Sentinel-2 based products. Moreover, we generally consider many more products compared to other studies (for instance, Mu et al. 2022 consider 6 30 m products; we consider over 13 across resolutions). We have explained this difference more clearly in the introduction on line 52:

Previous studies that have explored these discrepancies have focused on earlier-generation datasets that were generally coarser (~1 km), in line with the resolutions of the commonly deployed Earth observing satellites of the time, and did not include enough observations to provide time series of urban expansion[23,24]. More recent comparisons of higher resolution datasets, primarily in the remote sensing literature, are limited because they are either restricted to regional extents[25], or focus on comparing product accuracies, not area estimates and product typologies and their implications[19]. One recent study[26] examined the discrepancies between six 30 m global urban land products, though the main implication of these differences examined was for future urban projections. Here, we provide a comprehensive comparison of almost all (see Methods for selection criteria) the medium to fine resolution (100 m to 10 m) global urban datasets currently available (Table S1), showing that definitions of 'urban' remain a critical issue across these datasets, particularly in the newer 10 m resolution products. More importantly, we examine the consequences of the choice of dataset on a few common use cases relevant for examining urban climate change and its human impacts. For these use cases, the choice of dataset can have large influence on magnitude of urbanization impacts and, in some cases, even the direction of the urban climate signal.

More importantly, our goal was to move beyond dataset comparison and discuss, through actual analysis, what the choice of dataset means for specific use cases. Previous studies haven't focused on this application aspect of these datasets (beyond domain-specific case studies). We also analyze datasets used in the urban modeling literature and future projections of urbanization, which are both different from pure remote sensing products. As such, our study is much broader in scope than existing efforts. We have expanded on the relevance of this study on line 369:

In light of the development of multiple new global urban datasets at finer resolutions in the last decade, our goal here was to examine whether these state-of-the-art products provide better constraints on our understanding of urbanization across scales. The datasets considered here include thirteen historical urban land cover products (plus GHSL; see Methods), two datasets specifically used in process-based urban models, and four future projections of urban land. Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts.

(2) About the data products used in the manuscript, firstly, The GHSL products are also long-term series and relatively commonly used datasets. In addition, in my opinion, GlobeLand30 is also a very good datasets (https://www.un-spider.org/links-and-resources/data-sources/land-cover-map-globeland-30-ngcc). Therefore, this manuscript should compare the two datasets. From my research experience, the accuracy of 10-m resolution global urban land products, for example, ESA and ESRI land cover, is very low. But, in Table S2, their accuracy is still very high. Therefore, I think there is something wrong with your accuracy assessment. In this manuscript, you should give the absolute value of the area, not the percentage of impervious surface. For example, the total areas of global urban land.

There are a few separate points here that we respond to below:

1. Given the pace of dataset generation, we had some specific criteria when choosing datasets for inclusion, finally choosing, what we think, is a good mix of datasets from different groups, different methodology, and based on different data source. This is noted on line 452:

   We consider multiple global urban land cover datasets that have been developed over the last couple of decades to both examine differences between them across spatiotemporal scales and to discuss the impacts of these differences on a few use cases. Our focus here is primarily on datasets that have 100 m or finer resolutions, with the majority being derived from Landsat or Sentinel-2 satellite observations. We also consider the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover and European Space Agency Climate Change Initiative (ESA CCI) Land Cover datasets, which are at 500 m and 300 m, respectively. The former is one of the few physical estimates of global urban land that has been continuously updated since Potere et al. (2009)[24] and the latter because it is one of the few land cover products based on the Medium Resolution Imaging Spectrometer (MERIS) and has been used for multiple applications[27]. We do not consider any regional land cover datasets or land cover datasets released after 2021. This is why we did not focus on the latest version (P2023A) of the Global Human Settlement Layer (GHSL) in our main analysis.

2. In the original version of the manuscript, we did include GHSL, but only the 2018 data from the latest version, precisely because it has been commonly used in many applications, and because it was the only 10 m resolution estimate from GHSL (historical data are at coarser resolution since there was no Sentinel-2 before 2017). In the revised version, we have added both the time series for GHSL and the 2018 data and described the differences in more details in the methods on line 464:

   However, since this product has been used for various applications, we make an exception and provide results for 2018 ‘built spaces’ (at 10 m resolution) and the 5-year ‘degree of urbanization’ estimates from 1985 to 2020 (at 1000 m resolution) in the supplementary information (Fig. S3). The former is used for a small discussion about potential mismatches for urban applications (see Results section). For reference, the global urban ('built spaces') percentage in the 10 m GHSL P2023A is lower (0.49%) than the corresponding MODIS estimate (0.59%). Also note that the GHSL 10 m product includes road surfaces (global percentage of 0.02%) under a subset of 'open spaces', while 'built spaces' includes different types and heights of buildings.

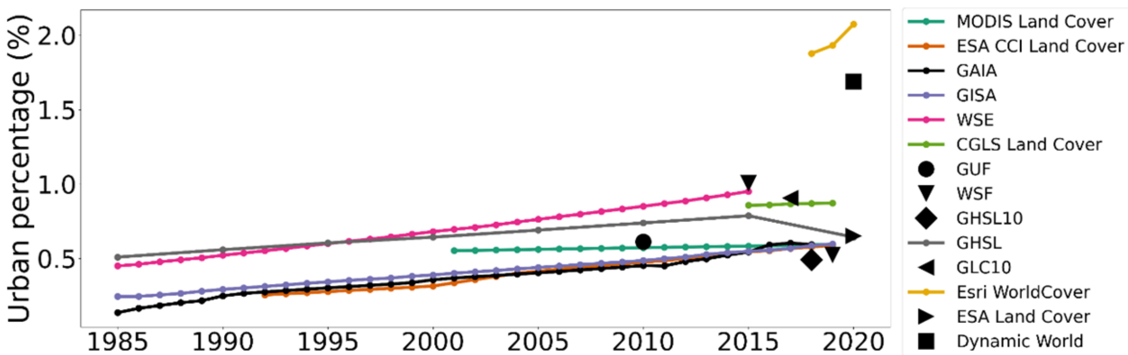
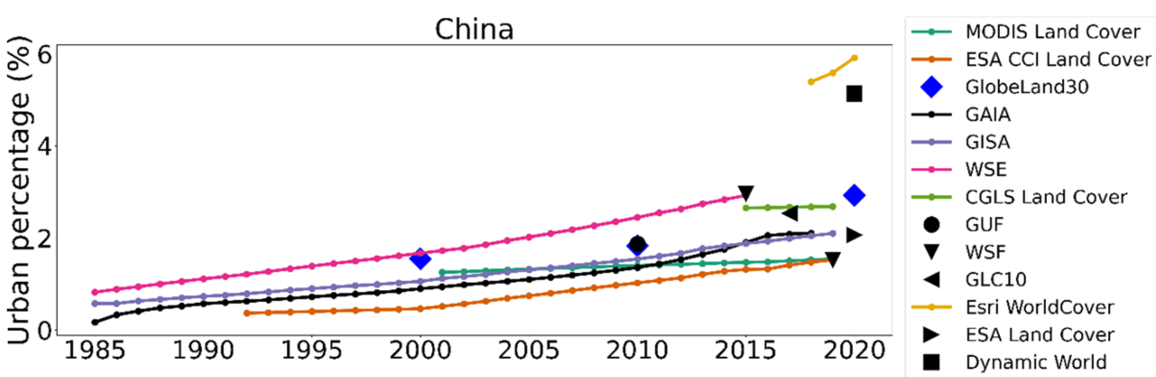


Fig. S3 | Urban percentage and its long-term changes across datasets (including GHSL) | Same as Fig. 4a of the main text, but includes the estimates of degree of urbanization ('Semi-dense urban cluster', 'Dense urban cluster', and 'Urban centre') every 5 years (from 1985 to 2020) and the 2018 10 m estimate of built spaces (labelled as GHSL10) from the latest version of the Global Human Settlement (GHSL) layer (P2023A).

3. Note that GlobeLand30 is difficult to access (the original website has been down for a while). Moreover, it is a somewhat old product, and we included a more recent product that is similar methodologically, namely FROM-GLC10 (we refer to it as just GLC10), in our analysis. Adding more datasets will not change our main results, since the primary goal of this study was not to do accuracy assessments. To illustrate for the benefit of the reviewer, we have calculated the urban percentage from GlobeLand30 for 2000, 2010, and 2020 for China, and compared to other datasets below. We have not added it to the manuscript since including this dataset does not add new insights to our conclusions.



4. We did a previous accuracy assessment of all land cover in the Dynamic World, ESA WorldCover, and Esri Land Cover data products (Venter et al. 2022) and found that,

across all three datasets, urban was the second most accurate class after water. As such, the low accuracy mentioned by the reviewer may be a regional scale issue. Our accuracy assessments here will also be biased high since these are overall accuracies since some of these products do not have other land cover classes. We want to reiterate that the accuracy assessment is not the goal of the paper. This is only done as a sanity check. We describe this in more detail on line 581:

Our primary goal in this study was not to focus on comprehensive accuracy assessments of these datasets. This is because of two main reasons. First, there have been multiple accuracy assessments of global land cover estimates across scales[16,18,22,24,25]. Second, given the differences in urban definitions in these datasets, standard accuracy estimates may not be particularly helpful. There has been discussion about the term 'ground-truth' in the broader remote sensing community that is relevant here[95]. However, as a sanity check, we provide basic accuracy estimates of the eight datasets used for representing present-day (2019/2020) urban land using the validation dataset created by the Dynamic World team[32]. The development of the training data employed 70 annotators, who manually labeled land use and land cover types in high-resolution images from Sentinel-2 for random dates in 2019. The annotation was done following the classification typology of the Dynamic World dataset. We chose this dataset since it is the largest available validation data that is relevant at the 10 m scale. Our accuracy estimates for the world and all the continents are summarized in Table S2 and show the percentage of the urban pixels in the reference that are correctly identified as urban in the dataset (overall accuracy). Based on this assessment, the GISA and WSF 2019 products perform the best followed by Esri Land Cover and Dynamic World. Overall, MODIS Land Cover performs the worst. However, beyond this sanity check, we should be cautious about the implications of these accuracy estimates. As an example, note that the urban definition in the Dynamic World dataset includes a mixture of residential buildings, streets, lawns, trees, isolated residential structures or buildings surrounded by vegetative land cover (Table S1). Therefore, it is a mosaiced land cover definition, partly because it uses a minimum mapping unit of 50 x 50 m. Other products like World Cover with 10 m x 10 m or GISA with 30 m x 30 m minimum mapping units may not be directly comparable (Table S1). Another example of a typological difference is that WSF 2019 does not even include roads. Basically, due to mismatch between land cover typology in the global datasets and the typology considered when creating ground truth data, it is difficult to provide an unbiased conclusion about relative accuracies of these datasets.

5. We provide percentage urban area since it is more intuitive to understand the degree of urbanization across regions for a variety of audience compared to absolute urban area values, which are in millions of square kms. In the revised manuscript, we also add a supplementary figure with those absolute values and discuss why we do not focus on that in the manuscript on line 182:

This pattern of rapid urban expansion is consistently echoed across all continents and for both absolute area (Fig. S1) and urban percentage (Figs 4, S2a; Fig. S2 is for the nations not in the continents examined in this main text), with Africa, Asia, and South America showing a notable rise (three- product means of 226.2%, 425.3%, and 186.5%, respectively), although from a lower baseline compared to North America and Europe. Note that we primarily focus on urban percentage throughout this analysis and the manuscript, rather than the absolute area, since the former is more intuitive for a wide range of audiences and allows easy comparisons for the extent of urbanization between regions with wide variations in area.

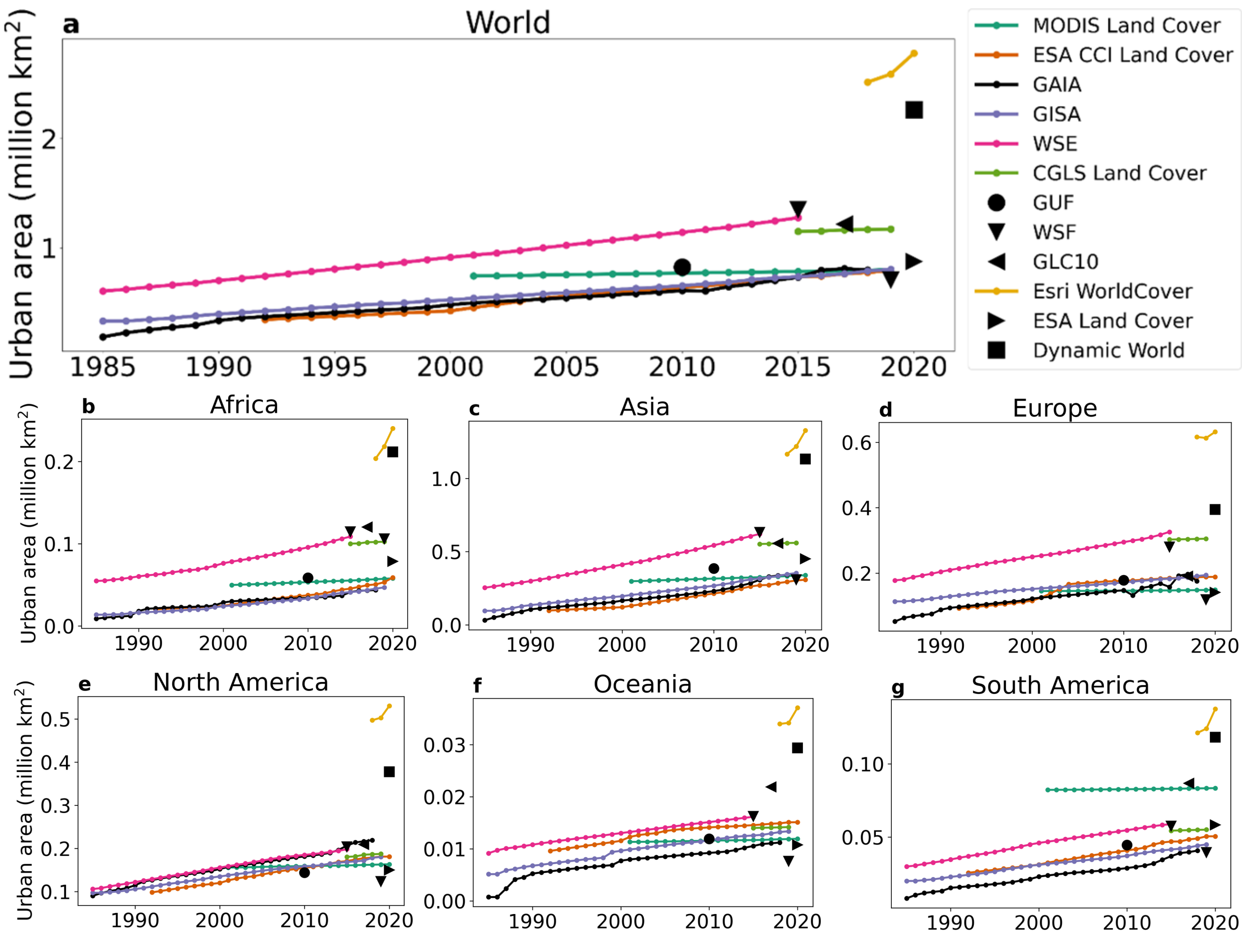


Fig. S1 | Urban area and its long-term changes across datasets | Urban area from 12 global data products for **a** World, **b** Africa, **c** Asia, **d** Europe, **e** North America, **f** Oceania, and **g** South America. Long-term changes are shown for datasets that span multiple successive years.

Note that the datasets that we are releasing with the paper have those absolute values, and readers can access them for further analysis.

We also add the total area of urban land on line 80, where we mention the urban percentage, for interested readers:

The eight-product mean global urban percentage is ~0.95% (1.27 million km^2).

(3) While the manuscript undertakes a commendable effort in comparing various global urban land datasets, it appears to miss a crucial opportunity to directly address and enhance

the practical application and customization practices already prevalent among urban researchers. For instance, it is common for researchers to adapt datasets like WSF2019 by integrating additional layers such as roads from OpenStreetMap, acknowledging that WSF2019 itself incorporates such data for SVM classification training. Furthermore, the practice of applying definitional adjustments to include urban green spaces or public areas, thereby transforming datasets like ESA and WSF for specific research needs, is well-documented. However, the manuscript does not seem to delve into how its findings could offer new insights or breakthroughs for these existing approaches. The potential of the study to guide researchers in navigating the complexities of dataset customization and to provide actionable insights for handling specific urban research challenges remains unclear. This gap suggests a missed opportunity to contribute to the ongoing discourse on the practical application of urban land datasets, potentially limiting the manuscript's relevance and utility for the broader community of urban researchers looking for novel methodologies and deeper understandings to enhance their data processing and application strategies.

Thank you for this comment. In the original manuscript, we did provide recommendations on various use cases and how they may be relevant to the physical systems being studied and for the nature of applications. Our examination of real uses of these data products were not merely hypothetical, but based on existing studies, which are cited in those sections. For instance, here is our discussion related to use of timeseries of urban land data to quantify flood risk (line 259):

Long-term changes in urban land are often combined with ancillary datasets to examine land use/ land cover transitions[35] and exposure to environmental hazards over time[33,11,8,9]. The rates of change over time would depend on the choice of dataset (Figs 4, 5), whereas most studies typically use a single product. For instance, ESA CCI Land Cover, GHSL, MODIS Land Cover, WSE have been individually used in these types of studies[33,11,8,9]. Andreadis et al. (2022)[33] and Rentschler et al. (2023)[8] both examined increased urbanization in flood-prone areas using two different urban land cover datasets (GAUD and WSE, respectively); therefore finding different magnitudes of these changes. We replicate a comparison of urban growth in flood plains[51] between 1985 and 2015 based on GAIA, WSE, and GISA here (Fig. 6b), with particularly large differences seen for Asia and Oceania. These larger differences reflect the stronger urban growth in the GAIA dataset (compared to the other two) in these continents (Figs 4c, 4f), with urban area increasing by 1395.6% and 998.5% between 1985 and 2015 in this product for Oceania and Asia, respectively, compared to 244.1% and 332.2% increase seen in the GISA dataset. The differences in urban growth between products for the other continents are not as large (303.4% and 476.2%,161.2% and 303.9%, and 173.5% and 229.4% for GISA and GAIA for Africa, Europe, and North America, respectively).

Another discussion on potential mismatches when using these land cover datasets for surface urban heat island mapping is on line 274:

Sometimes the choice of dataset can lead to artifacts due to mismatch between two products. For instance, Mentaschi et al. (2022)[9] combined the GHSL 2018 dataset with the MYD11 LST product to estimate intra-urban SUHI extremes. However, as noted earlier, this LST product is constrained by the MODIS Land Cover through the classification-based emissivity method[46]. As such, we would expect artifacts in LST for a proportion of pixels due to the GHSL data considering a pixel as urban while the emissivity in the LST product being defined for a rural surface (and vice versa). Similar artifacts would be expected for other combinations of MODIS LST with non-MODIS urban land cover estimates[52,53] or when MODIS LST is used to validate simulations from models that use different urban emissivity constraints[47,50,54].

Another discussion covers use cases for integration of urban datasets in process-based models (line 308):

As such, there are potential mismatches here that should be kept in mind. Since CESM currently does not use the low-density urban class within the urban model[66], it is implicitly assumed that anything up to medium-density, in terms of population, would be an appropriate representation of physical urbanization at the grid-scale. However, we find massive overestimations of urban land in CESM for regions of the world with high population density and low physical urbanization, such as Asia and Africa, with mean percentage errors of 179.9% and 136.2% (compared to GISA), respectively, for those continents (Fig. 7b). In contrast, this dataset largely underestimates urban land for North America and Oceania compared to GISA (mean percentage errors of -58.3% and -32.7%, respectively; Fig. 7b). Consequently, since urban surfaces in the CESM Land Model (CLM) uses specific prescribed thermodynamic, radiative, and morphological properties, if CLM has more urban land than the 'reality' (globally and for specific continents), urbanization would have a stronger bulk impact on grid-averaged surface climate variables in those regions, all else remaining equal. Exaggerated urban climate signals, such as urban heat and dry islands[67], should also be seen at the local scale, such as over the urban core, if the Jackson et al. (2010)[14] dataset is used in regional coupled simulations[64] that can resolve intra-urban advection. Similarly, since the MODIS Land Cover can be as low as 30% impervious, a model (like WRF) using MODIS as the land cover constraint would overestimate urban impacts on climate[2,54] for urban-to-rural transition zones, especially since urban models traditionally have not accounted for urban vegetation[55], which generally increases at the edges of cities.

It is not within the scope of the study to look at all possible use cases and adjustments. However, we do provide some recommendations, which we have expanded upon in the Discussions of the revised manuscript on line 369:

In light of the development of multiple new global urban datasets at finer resolutions in the last decade, our goal here was to examine whether these state-of-the-art products provide better constraints on our understanding of urbanization across scales. The datasets considered here include thirteen historical urban land cover products (plus GHSL; see Methods), two datasets specifically used in process-based urban models, and four future projections of urban land. Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts. We find large disagreements between the global urban data products across spatiotemporal scales. In fact, the largest divergences between datasets are seen for the most recent years from global to continental to country scales (Figs 4, 5) on inclusion of the new 10 m resolution products. At this resolution, it is possible to partially resolve urban vegetation, settlements, and roads, making the different urban definitions produce larger variations, with the Esri Land Cover and WSF 2019 datasets showing the highest and lowest urban land percentage, respectively (Fig. 1b). This variability in urban estimates across scales underscores the challenge in achieving a standardized measure of urban land even with globally available satellite observations.

We discuss the implications of these observed differences between the products for several use cases (Figs 6, 7, 8). The results demonstrate that it is important to be cognizant of the specifics of the datasets and any potential dependencies to ensure application-appropriate analyses. For instance, given the capability to separate roads and building roofs with current high-resolution satellite observations, particularly using Sentinel-2 and commercial satellites, we should rethink whether it makes physical sense to still include rooftops in estimates of many urban environmental risks, such as from floods and heat, given where urban residents are more likely to be exposed. Approaches to doing this may involve combining datasets with and without roads, ideally with otherwise similar urban typologies or adjusting datasets using masks for building rooftops. However, note that these approaches, and the data sources chosen, may have similar uncertainties, which one should be conscious about when exploring application workflows[73]. Note that there are other limitations to physiologically-relevant urban heat risk estimates using satellites that are beyond the scope of this study[67,74]. Similarly, as the newer urban land cover datasets are incorporated into process-based models, it is important to be aware of consistency in definitions between urban representations in models and the classification typologies used in data products. Using Dynamic World, which includes some vegetation in the urban class, is not appropriate if the model treats the entire urban grid as an impervious surface, such as in CESM, its offshoots[62,63], and most versions of WRF. Similarly, since WSF 2019 removes urban roads, incorporating this dataset into process-based models with a typical urban canyon structure, which includes both buildings and roads, is not recommended as doing so will capture only a fraction of the physical impact of urbanization on weather and climate. Structural differences between models are commonplace in the Earth sciences, which has encouraged the use of ensemble estimates to lend robustness to projections and for uncertainty quantification. Surface datasets, such as those for urban land, are an additional free parameter that can be largely decoupled from implementations of model physics. Uncertainty estimates due to differences in land use projections are much rarer[75], and do not currently resolve urbanization at finer scales[76]. Based on the differences seen here (Fig. 8), we recommend using multiple datasets, when possible and when the definitions of 'urban' align with the assumptions of the use case, both to provide more robust estimates of uncertainties for urban-resolving climate projections and to better quantify risks for rapidly urbanizing populations. In summary, the appropriate preprocessing methods of urban land products vary between datasets and for different applications. These should be determined case by case with guidance from relevant domain experts.

Minor revisions are as follows:

(1) This sentence 'there are currently at least four ~10 m resolution global land use land cover products, which include urban classes, and several urban-specific datasets that span multiple decades' is not correct, the earliest 10-meter global urban land product was in 2015.

There seems to be some confusion here. That sentence does not mean that all four datasets were developed recently. Rather, what we meant was, as of writing this manuscript, at least four such datasets exist. Moreover, as far as we know, the earliest global land cover dataset at 10-m resolution is the FROM-GLC10 data, which was officially published in 2019 (Gong et al. 2019). This dataset is for the year 2017 (uses Sentinel-2 images), but was trained using a sample collected for classifying Landsat data in 2015. That does not mean that the dataset was released in 2015. Note that we already examine the urban percentage in FROM-GLC10 in our manuscript. If the reviewer is talking about some other datasets that we have missed, we would be happy to refer to it. Additionally, we clarified that sentence on line 48:

There are currently at least four 10 m resolution global land use land cover products, which include urban classes, the earliest of which having been released in 2019[21], and several urban-specific datasets that span multiple decades[16,22].

(2) In the figure 3, I recommend using the absolute value of the area rather than the percentage of impervious surface.

We have added the absolute urban area plot for Fig. 3 (Fig. 4 in the revised manuscript) in the supplementary information (See below).

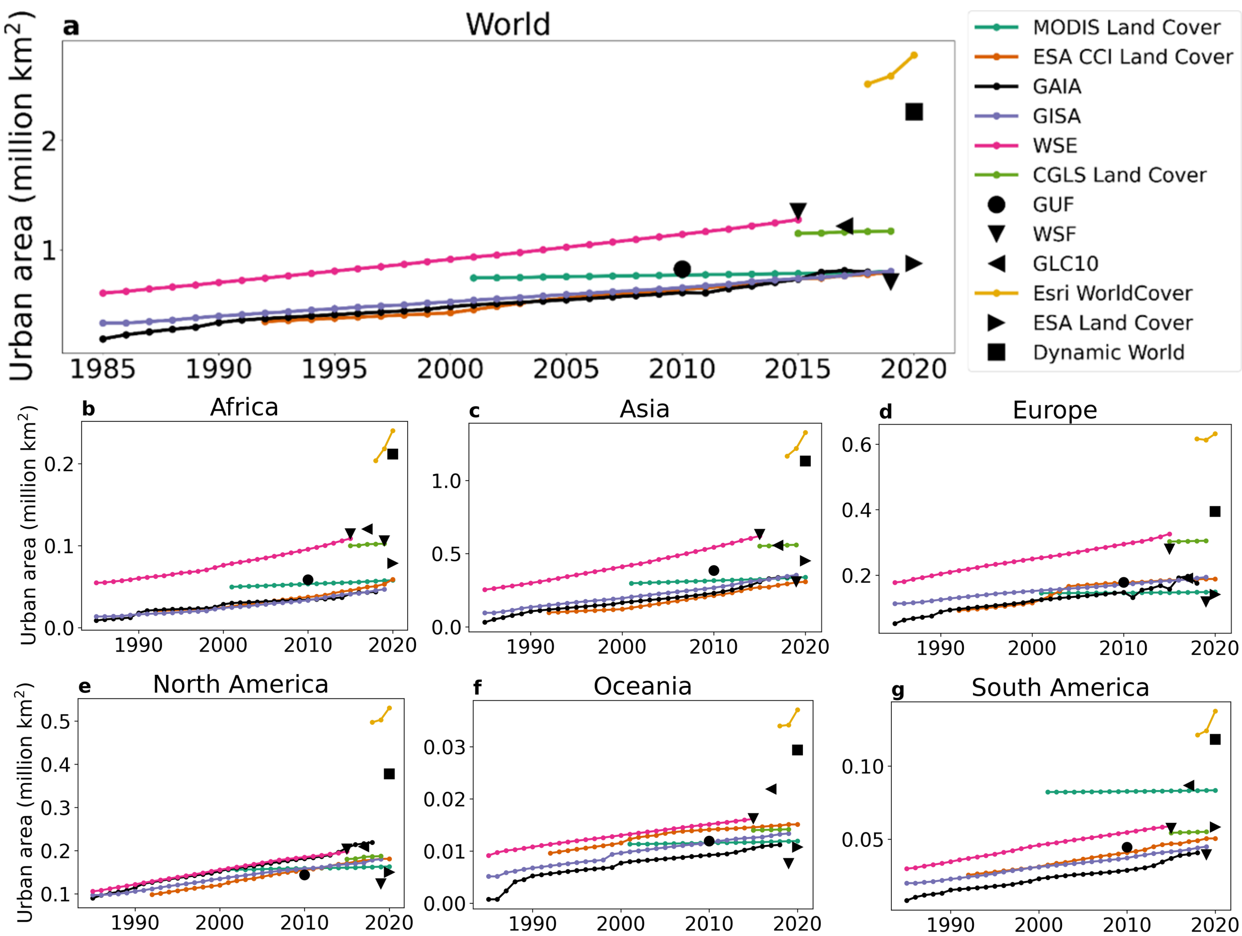


Fig. S1 | Urban area and its long-term changes across datasets | Urban area from 12 global data products for **a** World, **b** Africa, **c** Asia, **d** Europe, **e** North America, **f** Oceania, and **g** South America. Long-term changes are shown for datasets that span multiple successive years.

(3) It is recommended to add some references about other global urban land products, for example Hi-GISA (10-m resolution), Globeland30, GLC-FCS30D, and so on.

We have added these references at various points in the manuscript.

References:

1. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T., & Singh, G. (2022). Global 10 m land use land cover datasets: A comparison of dynamic world, world cover and esri land cover. *Remote Sensing*, *14*(16), 4101.
2. Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., ... & Song, L. (2019). Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Science Bulletin*, *64*(6), 370-373.
3. Mu, H., Li, X., Zhou, Y., Gong, P., Huang, J., Du, X., ... & Liu, D. (2022). Identifying discrepant regions in urban mapping from historical and projected global urban extents. *All Earth*, *34*(1), 167-178.

<u>Reviewer #2 (Remarks to the Author):</u>

The present study investigates a number of global urban land cover datasets and compare them w.r.t. to urban land cover fraction and long-term changes in the urban land cover. The authors have also discussed the implications of the differences between the datasets for practical examples and have produced a web tool. The analysis is comprehensive and well explained. Furthermore, the study covers an important topic in many different communities, especially for the (urban) climate modelling community. In particular, the different definitions of urban areas and the different resolutions of the data sets are the main reasons for the differences. With this analysis, the others provide guidance for different applications of the datasets. However, there are minor issues that need to be addressed before it can be published.

We thank the reviewer for the positive comments on this manuscript. In the revised version, we have made changes to address all of the minor issues.

Minor:
I like that the authors show examples on the impact of the differences on different practical applications. However, for one the example of urban land in flood plains there is only one sentence in the result section incl. one subfigure (Figure 4b). It would be interesting to explain the huge differences in Asia and Oceania.

We have expanded this portion in the revised version, starting with line 266:

We replicate a comparison of urban growth in flood plains[51] between 1985 and 2015 based on GAIA, WSE, and GISA here (Fig. 6b), with particularly large differences seen for Asia and Oceania. These larger differences reflect the stronger urban growth in the GAIA dataset (compared to the other two) in these continents (Figs 4c, 4f), with urban area increasing by 1395.6% and 998.5% between 1985 and 2015 in this product for Oceania and Asia, respectively, compared to 244.1% and 332.2% increase seen in the GISA dataset. The differences in urban growth between products for the other continents are not as large (303.4% and 476.2%,161.2% and 303.9%, and 173.5% and 229.4% for GISA and GAIA for Africa, Europe, and North America, respectively).

Line 56: It would be good to shortly introduce the selection criteria for the datasets her. Otherwise, the reader is wondering what “almost all datasets” mean.

Our selection criteria being somewhat complicated, we have added an explicit reference to the Methods on line 60:

Here, we provide a comprehensive comparison of almost all (see Methods for selection criteria)…

Lines 69ff: use 1 km2 instead of sq. km

Done.

Line 73: Esri Land Cover

Fixed.

Lines 81-83: Please remove "etc." since it is not clear, which countries are meant with "etc."

Done.

Lines 86-88: You are talking about urban growth here but the analysis of urban growth comes later in the manuscript. This can confuse the reader.

We have added a line here to refer users to the next section related to urban growth on line 100:

While all four of these regions are heavily urbanized, in the last few decades, the first two have shown stable urbanization levels, and the latter two have shown significant urban growth (see next section for analyses of urban growth).

Lines 91ff: You talk about "global to local scales" and "large scale". However, it is not really clear, what the difference between global and large scale. Also local scale can have a different meaning in different communities. Hence, it would be good to define these scales.

We have added the definitions of the scales on line 111:

The differences between these products are evident at all spatial scales, from global (for our entire planet) to national (by country) to the grids described above to local (for individual urban agglomerations).

Lines 242-246: This sentence is long and confusing. It is not clear, why there should be an overestimation if not all urban classes are used. It is also not clear, what is overestimated.

We meant urban land was overestimated. We have clarified this in the revision and shortened the sentence on line 309:

Since CESM currently does not use the low-density urban class within the urban model[66], it is implicitly assumed that anything up to medium-density, in terms of population, would be an appropriate representation of physical urbanization at the grid-scale. However, we find massive overestimations of urban land in CESM for regions of the world with high population density and low physical urbanization, such as Asia and Africa, with mean percentage errors of 179.9% and 136.2% (compared to GISA), respectively, for those continents (Fig. 7b). In contrast, this dataset largely underestimates urban land for North America and Oceania compared to GISA (mean percentage errors of -58.3% and -32.7%, respectively; Fig. 7b). Consequently, since urban surfaces in the CESM Land Model (CLM) uses specific prescribed thermodynamic, radiative, and morphological properties, if CLM has more urban land than the ‘reality’ (globally and for specific continents), urbanization would have a stronger bulk impact on grid-averaged surface climate variables in those regions, all else remaining equal.

Line 253: If roads are removed the assumption for some urban parameterization is not valid. Some assume an urban canyon, with buildings and roads. Hence, the recommendation should avoid such datasets for particular urban parameterizations.

We have updated this line 405:

Similarly, since WSF 2019 removes urban roads, incorporating this dataset into process-based models with a typical urban canyon structure, which includes both buildings and roads, is not recommended as doing so will capture only a fraction of the physical impact of urbanization on weather and climate.

282-300: While I think that future urbanization is an important topic, I am wondering why a new analysis is conducted in the discussion section. One could consider this as an outlook. It could also be listed as another application.

This is a good point. We have restructured the results and discussion such that the future urban projections are under applications. See line 336:

4) Future urban projections: Fourth, the differences in present-day and historical estimates of urban land also influence future urban projections. Several products have recently been developed to represent future urbanization scenarios[68–71] that can be used directly or incorporated into weather and climate models[12,56,72]. This use case also represents a special subset of the first point on derived datasets. We see large differences between these future estimates (Fig. 8a), which would depend on methodology (different growth models), input data (choice of historical urbanization estimate for model calibration and future population projection constraints), and assumed scenario of urbanization. For instance, Gao & O'Neill (2020)[69] consider distinct urbanization patterns across 375 sub-regions, while Chen et al. (2020)[68] use only 32 regions. Although both datasets are trained using GHSL, the Chen et al. (2020)[68] data are further calibrated against the ESA CCI estimate for 2015. The Li et al. (2021)[70] and He et al. (2023)[71] datasets are trained using annual nighttime light observations and the GAIA data, respectively. When we separate the sets of urban projections by continent, it becomes clear that the differences between them are not merely in terms of magnitude (Fig. 8). For instance, for the Shared Socioeconomic Pathways (SSP) corresponding to sustainable development (SSP1), the Li et al. (2021)[70] dataset shows most urban land at the end of the century for Africa, Asia, Europe, and South America. However, for the same scenario, He et al. (2023)[71] dataset shows the highest end-of-century urban percentage for North America, while the Chen et al. (2020)[68] dataset shows the highest urban land among datasets for Oceania. Another discrepancy between datasets is the gradual acceleration of urbanization in Gao & O'Neill (2020)[69] past mid-century versus a strong plateauing of urban growth in the He et al. (2023)[71] dataset. This discrepancy is related to the choice of future population projection used to develop the two datasets. These differences and inconsistencies between projections are seen for almost all the SSP scenarios, which reinforces the importance of being cognizant of dataset choice when reporting scientific results, quantifying impacts, and informing policy.

Caption Figure 4b: The sub-figure does not show correlations but scatterplot/linear regressions.

Changed to linear regressions.

Line 334: "one of the most"

Updated.

Line 396: Are there also negative values? Maybe you mean that both datasets need to have non-zero values?

That's correct. Changed to non-zero.

Figures 3d and S6: Previous studies also showed that there is a rapid increase rapid urban expansion in Europe in the ESA-CCI LC dataset (e.g. Li et al. 2021, Hoffmann et al. 2023), which is not shown in the other datasets. This increase can be also seen in the global urbanization rates (Li et al. 2018). Is there an explanation for this behavior? In any case, it would be good to mention this behavior in the manuscript.

We had noted this issue, but had avoided discussing it due to word limit. We have updated the text on line 207:

Another anomaly in the time series of urban percentage is seen over Europe for the ESA CCI product, with rapid expansion between 2000 and 2006 (Fig. 4d). This issue with the ESA CCI dataset has also been noted in other studies[43,44] and is related to a change in input dataset from the Medium Resolution Imaging Spectrometer (MERIS) baseline[27] to one with a resolution coarser than 300 m before 2003.

Web-tool: This is really interesting and helpful. However, I found that for MODIS tilted grids are show e.g. in Texas.

Thanks for catching this. The MODIS data used a different projection (does not impact the calculations, only visualization).

We have reprojected the image and updated the web app: https://ee-tc25.projects.earthengine.app/view/urbancomparison

Li, W., MacBean, N., Ciais, P., Defourny, P., et al. (2018): Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015), Earth Syst. Sci. Data, 10, 219–234, https://doi.org/10.5194/essd-10-219-2018

Radwan, T.M., Blackburn, G.A., Whyatt, J.D. et al. (2021) Global land cover trajectories and transitions. Sci Rep 11, 12814. https://doi.org/10.1038/s41598-021-92256-2
Hoffmann P, Reinhart V, Rechid D, de Noblet-Ducoudré N, et al. (2023): High-resolution land use and land cover dataset for regional climate modelling: Historical and future changes in Europe. Earth Syst. Sci. Data, 15, 3819–3852, doi: 10.5194/essd-15-3819-2023

Reviewer #3 (Remarks to the Author):

The manuscript addresses an increasingly important topic. The proliferation of global datasets, enabled by the decreasing computational costs and increasing availability of remotely sensed imagery at higher spatial resolutions and temporal frequencies, brings up the questions of to what extent these datasets indeed represent what they are claimed to represent and the degree of agreement among them. This is also one of the best written manuscripts I have reviewed for a long while. However, I would also like the authors acknowledge another, earlier work on comparing global urban land cover datasets, and relate their approach and findings to those of that study. Specifically, I would like to see a broad comparison and discussion of the another recently published study (Mu et al. 2022) not cited in the current manuscript. I would like to know in what ways the authors work is similar or not to that one with an emphasis on where these two studies agree (or not). Other than that, I have a few suggestions and comments to help improve the clarity and flow of the paper (please see the annotated manuscript and supplementary file).

Reference:
Mu, H., C. Xu, D. Liu, X. Li, Y. Zhou, P. Gong, J. Huang, X. Du, J. Guo, W. Cao and Z. Sun (2022). "Identifying discrepant regions in urban mapping from historical and projected global urban extents." All Earth 34(1): 167–178.

Thank you for the positive comments on the manuscript. In the revised version, we have addressed all the concerns. We have added a discussion on the Mu et al. 2022 paper on line 58:

One recent study[26] examined the discrepancies between six 30 m global urban land products, though the main implication of these differences examined was for future urban projections.

We have also done new analysis to replicate the methodology in Mu et al. to compare our results. This is on line 105:

We also calculate the coefficient of variation between the eight present-day estimates for 0.9° x 1.25° grids over the Earth's surface (Fig. 2b). This approach is similar to that of Mu et al. (2022)[26] and avoids the variabilities in how regions should be defined beyond geopolitical boundaries. Although we use a different (partially overlapping) set of urban data products compared to the Mu et al. (2022)[26] study, there are some common regions with disagreements between datasets in both studies, namely Southeast and Central Asia. We also find large disagreements in East Africa (Fig. 2b).

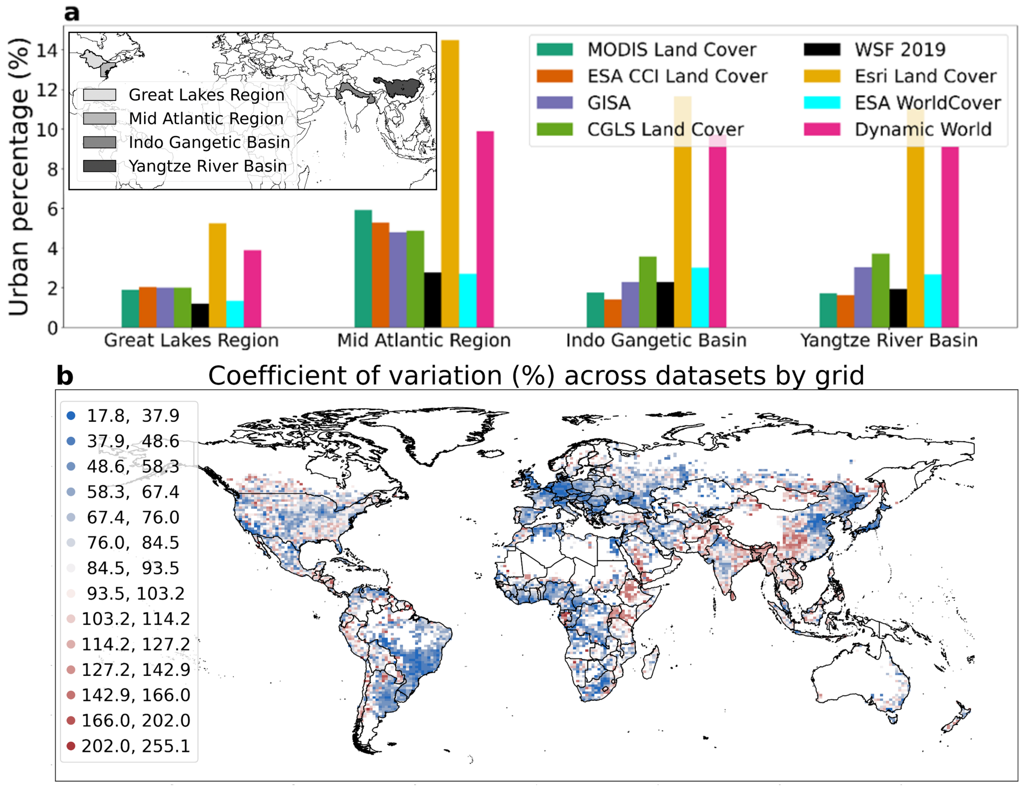


Fig. 2 | Present-day urban estimates across datasets from regional to grid scale | Sub-figure **a** shows the urban percentage across eight datasets for four selected regions in the world. The extent and location of these regions is shown in the inset. Sub-figure **b** shows the coefficient of variation (standard deviation divided by mean) among those urban estimates for 0.9° x 1.25° grids. The legend value ranges exclude the upper bound.

And discussed how our study differs from existing studies that are more focused on the accuracy assessment from the remote sensing perspective on line 369:

In light of the development of multiple new global urban datasets at finer resolutions in the last decade, our goal here was to examine whether these state-of-the-art products provide better constraints on our understanding of urbanization across scales. The datasets considered here include thirteen historical urban land cover products (plus GHSL; see Methods), two datasets specifically used in process-based urban models, and four future projections of urban land. Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts.

Below are the responses to the annotations in the manuscript and supplementary file.

On main text:

Not only natural land? Croplands are not "natural" strictly speaking.

We have updated this on line 31:

Urbanization, the global shift of rural to urban societies, leads to replacement of natural land, as well as cropland, with roads, buildings, pavement, parks, etc[1].

And surely there are more suitable sources to cite than number 2 here.

That location was referring to two different ideas, one on the definition of the urbanization and another on the climate impacts. The second paper is a review of these impacts. We have now separated these two lines and added another citation for the impacts in addition to the previous one on line 31:

Urbanization, the global shift of rural to urban societies, leads to replacement of natural land, as well as cropland, with roads, buildings, pavement, parks, etc[1]. These land use/land cover transitions, combined with anthropogenic activities in cities, together impact local energy, water, and carbon budgets[2,3].

I suggest the authors first mention the countries with the most urban land and then those with the highest percentages of urban land. For many reasons (the number of people affected, the scale of impact on climate etc), the urban land cover and its change over time in a country like China is more significant than the very high per cent urban land cover in a tiny country such as the Vatican city (which is really a part of the wider Rome metro area from a pure physical urban land cover perspective).

We have restructured the text on line 71 according to this suggestion:

China is the country with the most urban land (264403 km^2 covering 2.82% of its total area based on the eight-product mean), followed by the United States (183735 km^2; 1.94%), India (85760 km^2; 2.77%), and Russia (59311 km^2; 0.35%) (Fig. 1b). Overall, Vatican City and Singapore show the highest percentages of urbanization (eight-product mean of 79.87% and 53.7%, respectively). Ignoring uninhabited territories, on the low end, there are several overseas island territories, such as Cocos, Midway, and Pitcairn Islands, with negligible urban land detected by these satellite-derived products.

This is a good idea. However, the authors should also provide a more complete list of areas where there are notable differences among the datasets (in a table, perhaps in supplementary) beyond these two examples (Shanghai and Delhi).

We show Shanghai and Delhi as two examples of local-scale urbanization and provide the web app to allow users to explore other such examples. Since there are many possible ways to define regions, it would be unfeasible to provide a complete list of such areas, other than the country-level estimates we already provide (and linked through Figshare). Instead, we performed a new analysis of disagreements by grid for the whole globe that more clearly illustrates regions with discrepancies between datasets. This is noted on line 105 (see earlier reply for relevant figure; Fig. 2b):

We also calculate the coefficient of variation between the eight present-day estimates for 0.9° x 1.25° grids over the Earth's surface (Fig. 2b). This approach is similar to that of Mu et al. (2022)[26] and avoids the variabilities in how regions should be defined beyond geopolitical boundaries. Although we use a different (partially overlapping) set of urban data products compared to the Mu et al. (2022)[26] study, there are some common regions with disagreements between datasets in both studies, namely Southeast and Central Asia. We also find large disagreements in East Africa (Fig. 2b).

On Supplementary file:

This needs explanation. According to the U.N. regions (which is 'official'), the U.K and Ireland are included in Western Europe; Greenland and the Bahamas are included in the North America region. But more importantly, why? Where this figure is cited, there is no mention of islands so what is the point?

We use the World Bank national boundaries, which include a handful of rows not within the other main continents; we focus on the main continents in the main text. Note that these are not places like Greenland, U.K., etc. We added that additional plot to the supplementary file since we did not want to leave out any group of nations from our analysis. We have clarified this on line 182:

This pattern of rapid urban expansion is consistently echoed across all continents and for both absolute area (Fig. S1) and urban percentage (Figs 4, S2a; Fig. S2 is for the nations not in the continents examined in this main text),..

We are happy to remove this supplementary figure if the reviewer thinks it's not useful.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you for taking part in this important initiative. We have addressed all the concerns from your co-reviewer above.

REVIEWERS' COMMENTS

Reviewer #3 (Remarks to the Author):

The authors did a good job in addressing all the reviewers' comments.

Reviewer #5 (Remarks to the Author):

This study compared the mainstream medium to fine-resolution global urban datasets with spatial resolution ranging from 10 to 100 m, and found that the large discrepancies in the definition of 'urban' and derived estimates in spatiotemporal urbanization, presented as statistics. The topic is interesting and useful. However, my main concern is the limited novelty and advanced knowledge in the field. The main body is about data comparison.

For example, the key insights such as "there are substantial discrepancies in urban land area estimates among them influenced by scale, differing urban definitions, and methodologies" is not new. For any new urban/impervious area/human settlement mapping datasets generated, the associated paper will quantitatively compare with some of the state-of-the-art benchmark datasets, the findings are similar. Although the authors compared relatively larger sets of benchmark datasets, the novel insights are limited, especially for the interdisciplinary readership in Nature Communications. Moreover, the authors did not dive into taking some of the grand challenges, e.g., to what extent the definition of 'urban', the methods used, and the scale will impact the derived urbanization quantification. And how this kind of empirical evidence can be turned into underlying mechanisms to account for the reported urbanization discrepancy?

Additionally, by integrating case studies related to monitoring urban climate risks and modeling urbanization-induced impacts on weather and climate from regional to global scales, the authors mention that the results demonstrate the importance of choosing fit-for-purpose datasets for examining specific aspects of historical, present, and future urbanization. However, these case studies, along with the data comparison, further consolidate my concerns regarding the lack of in-depth scientific novelty.

Large disagreements in estimates of urban land across scales and their implications

(Original comments in **bold**, response in regular font, and manuscript revision in green)

Reviewer #5 (Remarks to the Author):

This study compared the mainstream medium to fine-resolution global urban datasets with spatial resolution ranging from 10 to 100 m, and found that the large discrepancies in the definition of 'urban' and derived estimates in spatiotemporal urbanization, presented as statistics. The topic is interesting and useful. However, my main concern is the limited novelty and advanced knowledge in the field. The main body is about data comparison.

We are glad that the reviewer has found the topic interesting and useful. We provide the point-by-point response to the comments below.

For example, the key insights such as "there are substantial discrepancies in urban land area estimates among them influenced by scale, differing urban definitions, and methodologies" is not new. For any new urban/impervious area/human settlement mapping datasets generated, the associated paper will quantitatively compare with some of the state-of-the-art benchmark datasets, the findings are similar. Although the authors compared relatively larger sets of benchmark datasets, the novel insights are limited, especially for the interdisciplinary readership in Nature Communications.

Thank you for the comments. It is correct that when researchers develop a new dataset, they compare against a set of state-of-the-art datasets. However, these comparisons are rarely as comprehensive as the analysis presented here. More importantly, we then use these different datasets to examine the role of choice of datasets on various use cases. Note that several of these use cases are often published in general interest journals like Nature Communications and even Nature. As such, showing the importance of choice of datasets is relevant for the broader community of users of these datasets, especially those who are do not have a background in satellite remote sensing and land use land cover mapping. In the previous version of the manuscript, we had mentioned the advances in this paper compared to previous studies on line 52:

Previous studies that have explored these discrepancies have focused on earlier-generation datasets that were generally coarser (~1 km), in line with the resolutions of the commonly deployed Earth observing satellites of the time, and did not include enough observations to provide time series of urban expansion[23,24]. More recent comparisons of higher resolution datasets, primarily in the remote sensing literature, are limited because they are either restricted to regional extents[25], or focus on comparing product accuracies, not area estimates and product typologies and their implications[19]. One recent study[26] examined the discrepancies between six 30 m global urban land products, though the main implication of these differences examined was for future urban projections. Here, we provide a comprehensive comparison of almost all (see Methods for selection criteria) the medium to fine resolution (100 m to 10 m) global urban datasets currently available (Table S1), showing that definitions of 'urban' remain a critical issue across these datasets, particularly in the newer 10 m resolution products. More importantly, we examine the consequences of the choice of dataset on a few common use cases relevant for examining urban climate change and its human impacts.

And on line 373:

Unlike past studies, primarily in the remote sensing literature, that focus on accuracy assessments[22,24,25] or compare newly developed data products against other available datasets[18,19,32,70], our main goal was to explore what these disagreements mean for applications of these datasets in modeling and observational studies. As such, this study is broader in scope compared to those efforts and relevant to researchers and practitioners interested in the urban environment beyond remote sensing experts.

Moreover, the authors did not dive into taking some of the grand challenges, e.g., to what extent the definition of 'urban', the methods used, and the scale will impact the derived urbanization quantification. And how this kind of empirical evidence can be turned into underlying mechanisms to account for the reported urbanization discrepancy?

In the previous version of the manuscript, we had explicitly provided several discussions and recommendations relevant to the use cases examined. For instance, on both observational and modeling estimates using these datasets on line 388:

We discuss the implications of these observed differences between the products for several use cases (Figs 6, 7, 8). The results demonstrate that it is important to be cognizant of the specifics of the datasets and any potential dependencies to ensure application-appropriate analyses. For instance, given the capability to separate roads and building roofs with current high-resolution satellite observations, particularly using Sentinel-2 and commercial satellites, we should rethink whether it makes physical sense to still include rooftops in estimates of many urban environmental risks, such as from floods and heat, given where urban residents are more likely to be exposed. Approaches to doing this may involve combining datasets with and without roads, ideally with otherwise similar urban typologies or adjusting datasets using masks for building rooftops. However, note that these approaches, and the data sources chosen, may have similar uncertainties, which one should be conscious about when exploring application workflows[73]. Note that there are other limitations to physiologically-relevant urban heat risk estimates using satellites that are beyond the scope of this study[67,74]. Similarly, as the newer urban land cover datasets are incorporated into process-based models, it is important to be aware of consistency in definitions between urban representations in models and the classification typologies used in data products. Using Dynamic World, which includes some vegetation in the urban class, is not appropriate if the model treats the entire urban grid as an impervious surface, such as in CESM, its offshoots[62,63], and most versions of WRF. Similarly, since WSF 2019 removes urban roads, incorporating this dataset into process-based models with a typical urban canyon structure, which includes both buildings and roads, is not recommended as doing so will capture only a fraction of the physical impact of urbanization on weather and climate.

In the revised manuscript, we have further added to these recommendations focusing on the scale of application on line 413:

Based on the differences seen here (Fig. 8), we recommend using multiple datasets, when possible and when the definitions of 'urban' align with the assumptions of the use case, both to provide more robust estimates of uncertainties for urban-resolving climate projections and to better quantify risks for rapidly urbanizing populations. Moreover, for local- to regional-scale science and applications, it is generally preferrable to use maps developed for those areas instead of global datasets, since the former are better calibrated to capture the unique urban development patterns of those areas. In summary, the appropriate preprocessing methods of urban land products vary between datasets and for different applications and scales. These should be determined case by case with guidance from relevant domain experts.

More importantly, we want to start a communication that helps both data users and data producers talk about these uncertainties more clearly, which we have further added to in this version. This is noted on line 443:

Given the variations in the temporal and spatial patterns of urban land across datasets seen here, we suggest that decision makers consider a variety of datasets to inform policies from urban to national scales. Similarly, data producers should try to be more up front about the assumptions underlying their datasets, particularly when the intent is to assist policymaking. In a world of cheap compute and increasingly available high-resolution satellite imagery, we expect such global data products to continue to be published. Moreover, these datasets will continue to be used for large-scale studies to answer various research questions[43,89,90], some quite outside the range of uses that have been illustrated in the present study. Given this rapid pace of dataset generation and usage, a concentrated and ongoing effort is needed from the urban scientific community to assess fit-for-purpose datasets for distinct science and applications before they are adopted to quantify the costs and benefits of environmental solutions and support broader policies.

Additionally, by integrating case studies related to monitoring urban climate risks and modeling urbanization-induced impacts on weather and climate from regional to global

scales, the authors mention that the results demonstrate the importance of choosing fit-for-purpose datasets for examining specific aspects of historical, present, and future urbanization. However, these case studies, along with the data comparison, further consolidate my concerns regarding the lack of in-depth scientific novelty.

It is unclear what the reviewer means by this. Note that the case studies were meant to provide examples that follow from the discrepancies in area seen in the dataset comparison. We do not plan to look at all possible use cases, since that is impossible to do in a single paper. In the revised manuscript, we have clarified this before the point about how similar sensitivity analyses and methodological considerations are critical for other global use cases. This is noted on line 448:

Moreover, these datasets will continue to be used for large-scale studies to answer various research questions[43,89,90], some quite outside the range of uses that have been illustrated in the present study. Given this rapid pace of dataset generation and usage, a concentrated and ongoing effort is needed from the urban scientific community to assess fit-for-purpose datasets for distinct science and applications before they are adopted to quantify the costs and benefits of environmental solutions and support broader policies.